Stat 462/862 Assignment 1

(Due on Sept 27, 2018, hard copy, in the class)

1. Use *data*(*state*) to load the data *state* in R. Consider one of its dataset, *state.x77*, and complete the following parts.

   (a) Look up the help for *state.x77*.

   (b) Compute the dimension size of *state.x77* and display its dimension names.

   (c) Write an expression that returns the summary statistics, as given by *summary()*, for each of the columns in this matrix.

   (d) Create a vector named *popdense* that contains the population density in persons per square mile. Note that the population in *state.x77* is given in thousands.

   (e) Apply the function *pairs()* to the matrix to create scatter plots of all columns against one another.

   (g) Select the states that have $< 1\%$ illiteracy, but a murder rate of > 10 per 100,000.

   (h) Select the states that have greater than average high school graduate rates, but less than average annual income.

   (i) Create a vector named *Murder* that contains the murder rates.

   (j) Create a vector named *Illiteracy* that contains the illiteracy rates.

   (k) Create a design matrix, X, that contains all 1's in the first column and the illiteracy rates in the second column. This will serve as our design matrix in part (m).

   (l) Assume that there is an approximate linear relationship between *Illiteracy* and *Murder rates*. Given the design matrix defined above, we can define a simple linear model as $y = X\beta + \epsilon$ where y(*Murder*) is the dependant variable, X is the design matrix of independent variables, β is the vector of parameters and ϵ is the error term. The least squares estimate of β is: $\hat{\beta} = (X^T X)^{-1} X^T y$, where X^T is the transpose of X. Write a function which has inputs X and y, and returns $\hat{\beta}$ in R.

   (n) In a single plot, draw the following subplots (1) a scatter plot of *Murder* (y) versus *Illiteracy* (x); (2) a histogram of *Murder*; (3) a Quantile-Quantiale plot of *Murder*;(4) a boxplot of *Murder* and *Illiteracy*.

	0.75	0.9	0.95	0.975	0.99	0.999
1	1.0000	3.0777	6.3138	12.7062	31.8205	318.3088
2	0.8165	1.8856	2.9200	4.3027	6.9646	22.3271
⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	0.6786	1.2958	1.6706	2.0003	2.3901	3.2317
Inf	0.6745	1.2816	1.6449	1.9600	2.3263	3.0902

2. (Graduate students only) Tabulating quantiles of the t-distribution.

   (a) Create a vector called *percentile* that contains the values 0.75, 0.9, 0.95, 0.975, and 0.99 and 0.999.

   (b) Create a vector named *df* that contains the integers 1 to 30 followed by 60 and Infinity. Do this without typing in the integers 1 through 30.

   (c) Create a matrix named *tTable* that returns the percentile from a t-distribution where the rows represent the degrees of freedom as specified by the vector *df* and the columns are the *quantiles* as specified by percentile. That is, the cells are t such that $P(T_{df} <= t) = percentile$ where T_{df} is a t-distribution with *df* degrees of freedom. Hint: you may need to perform a transpose to get the rows and columns as specified.

   (d) Round the contents of *tTable* to four decimal places.

   (e) Assign the values in *df* and *percentile* as row and column names respectively so that when the matrix is displayed the first and last couple of rows look like the following:

3. (Multiple linear regression) Install and load the R package $ISLR$ and consider the dataset *Auto* in the package. Treat *mpg* as the dependent variable and all the other variables except *name* as the independent variables (predictors). Note that the 1,2,3 of the variable *origin* correspond to $American, European, Japanese.$

   (a) Create a pairwise scatter plot for dependent and independent variables. Show the plot and make comment on the plot.

   (b) Computer the correlation matrix between the variables using $cor()$ function.

(c) Fit the multiple linear regression model. Show the table of the fitted model: coefficients estimation, their standard deviation, t-statistic, and p-values. Show R^2 and the estimation $\hat{\sigma}^2$.

(d) Obtain the prediction of mean response, its associated prediction error and $100(1-\alpha)\%$ confidence interval based on the fitted model for the new input $cylinders = 8, displacement = 300, horsepower = 150, weight = 3600, acceleration = 13, year = 77, origin = 3$.

(e) Is there relationship between the independent variables and the response variable?

(f) Which predictors appear to have a statistically significant relationship to the response?

(g) Produce the residuals plots.